Epistemic Assortativity in Scientific Collaboration Networks

Ignacio Toledo, Jorge Fábrega

Universidad del Desarrollo, Facultad de Gobierno, Centro de Investigación en Complejidad Social (CICS), Chile.

INTRODUCTION

Understanding how scientists collaborate during the production and publication of scientific knowledge has become an issue of great importance.

This phenomena has been thoroughly studied through the analysis of scientific co-authorship networks, where nodes are scientists and links are papers with shared authorship.

One major finding in co-authorship networks shows high levels of assortativity by degree, and high levels of clustering. This means more gregarious scientists tend to be connected to each other, and that two scientists have higher chances to collaborate if they share a common co-author.

However, these findings do not give a clear insight into the individual motivations underpinning the observed macroscopic pattern in the structure of the collaboration networks. Such mechanisms are still under discussion. We address this issue considering the research interests compatibility as a key factor in the collaborator selection process.

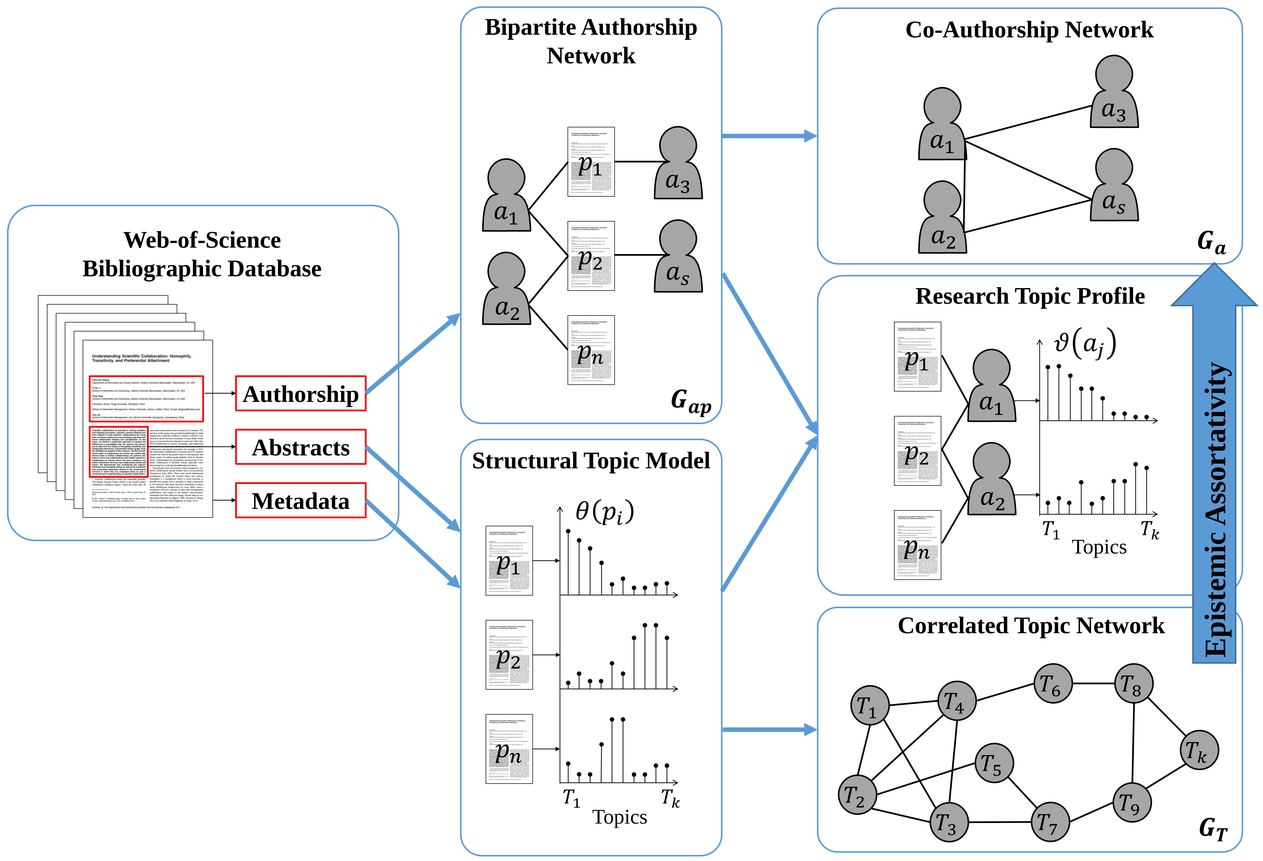
We propose a new method to represent the research interests of each author based on a topic model, estimated using the abstracts of the papers of a bibliographic database. We further test for assortativity by degree and by our research interest compatibility measure. The results show that both assortativity indexes are remarkably similar.

The results suggest that research interests compatibility plays a significant role in the link formation process in scientific collaboration networks.

METHODS

Bipartite Authorship Network, G_{ap}: This network is composed of two types of nodes: *authors* and *papers*. A link represents an authorship relationship.

Structural Topic modeling: by using the corpus of abstracts of the papers in the database, we estimate a probabilistic topic model, using a natural language processing (NLP) technique called Structural Topic Model (STM), which allow us to add papers metadata as covariates. The model output is composed of a probability distribution over words for each topic k, and a probability distribution over topics $\theta(p_i)$ for each document p_i.



Co-Authorship Network, G_a: we perform the projection of the bipartite authorship network over the authors layer. Nodes are authors and links are papers with shared authorship.

Research Topic Profiles: by using the paper distributions over topics, combined with the authorship bipartite network (*author-paper*), we construct a topic profile for each scientist, that represents their research interest portfolio.

The research topic profile $\vartheta(a_j)$ of a scientist a_j is calculated as the mean topic distribution based on papers in which the scientist has collaborated, represented by $\Gamma(a_j)$.

$$\vartheta(a_j) = \frac{1}{|\Gamma(a_j)|} \sum_{p_i=\Gamma(a_j)} \theta(p_i)$$

This means that a scientist has a proportion value for each topic in the range of 0 to 1, and that the sum of all these proportion values equals one.

Correlated Topic Network, G_T: we also construct a semantic network of topics based on that topics are distributions over words. In this network, two topics are connected if the correlation factor between their distributions is greater than an arbitrary threshold.

Epistemic Assortativity: we perform an assortativity test by the product of proportions of related research topics. This measure represents the interest of scientists in subfields of knowledge composed of two research topics related semantically.

The proposed measure is calculated for each edge $e = \langle T_m, T_n \rangle \in E(\boldsymbol{G_T})$, as follows:

$$R_{e,j} = \vartheta(a_j, T_m) \cdot \vartheta(a_j, T_n), \forall j \in [1, s]$$

This measure feeds the assortativity coefficient r_e for the edge e. The final coefficient r is the average value of r_e.

$$r = mean(r_e) = mean\big(Assortativity(\boldsymbol{G_a}, R_e)\big)$$

DATASET

We use three bibliographic databases obtained from Web-of-Science for Economics, Sociology and Neuroscience. Each contains abstracts and rich metadata, of papers, from a 13 year period (2004-2017).

RESULTS

The assortativity coefficients obtained from the proposed methodology show that scientists interests on research topics play a significant role in the mutual selection process of authors to collaborate with when working in the production of knowledge. Moreover, the magnitude of the coefficient r_{R_e} is comparable with assortativity by degree r_{degree}. With the exception of Neuroscience, in both Economics and Sociology r_{R_e} is slightly greater than r_{degree}.

Table I: Assortativity Coefficients

Discipline	r_{degree}	r_{R_e}
Economics	0.688	0.715
Sociology	0.441	0.477
Neuroscience	0.909	0.697

DISCUSSIONS

These results suggest that scientists are prone to collaborate with colleagues with similar research interests nearly as much as with colleagues with a similar degree. Our results provide a possible explanation on the mechanism underlying the formation of co-authorship relations. These findings will help us develop more accurate link-prediction models in Scientific Collaboration Networks than models based on solely topologic measures. However, the relationship between research topics and degree requires a more detailed examination.

REFERENCES

- Newman, M. E. J. (2000). The structure of scientific collaboration networks.
- Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks.
- Newman, M. E. J. (2002). Mixing patterns in networks.
- Barabási et al. (2002). Evolution of the social network of scientific collaborations
- Blei et al. (2003). Latent Dirichlet Allocation.
- Roberts et al. (2013). The structural topic model and applied social science.
- Roberts et al.(2015). stm: R Package for Structural Topic Models. Journal of Statistical Software
- Liben-Nowel & Kleinberg (2007) The link prediction problem for social networks